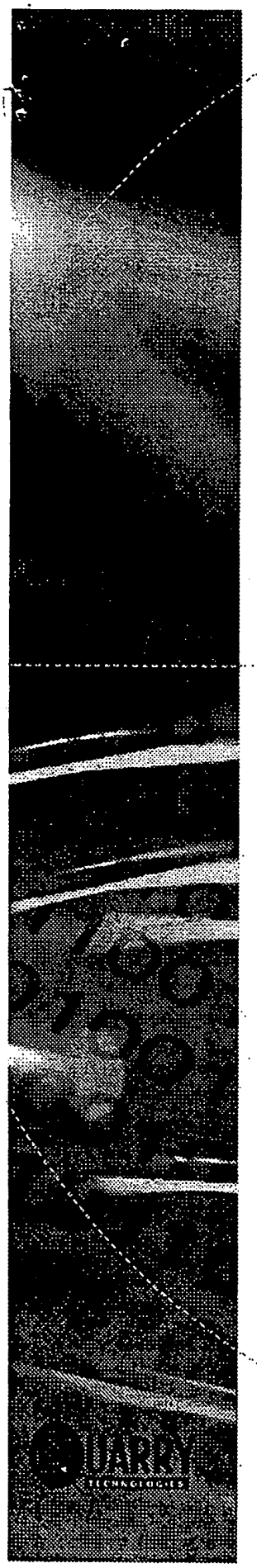

WHITEPAPER

Anatomy of an IP Service Edge Switch

Accelerating Advanced IP Services with a Pipelined Architecture

Steve Kohalmi, Chief Systems Architect
Richard Forberg, Vice President of Product Management

Anatomy of an IP Service Edge Switch

Accelerating Advanced IP Services with a Pipelined Architecture

Steve Kohalmi, Chief Systems Architect
Richard Forberg, Vice President of Product Management

Advances in technology are increasing bandwidth both in the core of the network as well as at the access points. But all this bandwidth will not pay off for the Network Service Providers (NSPs) unless high-value IP-based network services can be delivered cost effectively using an integrated platform. To accomplish this goal, NSPs are now turning to the next-generation IP service edge switch.

Traditional edge aggregation routers have done little more than provide basic IP packet forwarding. IP service platforms, on the other hand, can perform many diverse functions in addition to routing in the delivery of customized services to subscribers. The list of functions being demanded of these intelligent service delivery platforms is constantly growing, as NSPs seek to eliminate hard-to-manage Customer Premises Equipment (CPE) and generally consolidate platforms used in their Points of Presence (PoPs) with IP as the common foundation for all services.

"...this pipelined architecture...will set the course of future development in optical-speed IP service delivery."

First-generation IP service routers, though quick to market, were limited by their architecture. These devices placed a heavy reliance on general-purpose processors, often using many of them operating in parallel. This early architecture is now facing scalability and functionality limitations, as well as increased management complexity due to its many, slow, parallel processing paths. The initial platforms based on this design do not support processing of multiple services simultaneously, nor are they able to handle individual tunnels with high throughput, or any type of high-bandwidth subscriber connection. This makes them unsuitable for new optical access services or even high-speed leased line services.

The solution to this problem is the next-generation IP service edge switching architecture, which uses a hardware-based pipeline built on leading-edge network processors combined with high-speed ASICs and encryption co-processors. Yielding the maximum performance while enhancing feature flexibility, this pipelined architecture represents a dramatic improvement that will set the course of future development in optical-speed IP service delivery. It is a revolution comparable to when routing systems transformed from software-based routers with general purpose CPUs into ASIC-based Layer 3 switches.

PACKET PROCESSING STAGES FOR IP SERVICES

New requirements, standards, and protocols in the areas of Quality of Service (QoS), traffic engineering, and security have increased the complexity of packet processing. A sophisticated solution requires not only speed, but also fine granularity on the services delivered, to meet the unique business needs of all customers and their respective subscribers. For

support of this solution, the ability to both monitor and enforce the service levels delivered to each subscriber, while collecting all required accounting data, is essential.

To deliver this high degree of customization and manageability requires many processing stages, spanning many IP-related protocols, link-layer encapsulations, and tunneling technologies. (*See Figure 1.*) For the most advanced suites of IP services, a series of stages must be executed for every packet that enters the system from any given subscriber.

Stage 1 - Remove Link-layer Headers & Decrypt. In simple text-book IP networks there is only one link-layer header, such as an Ethernet or PPP frame. But, in the complex environment of today's access networks there can be multiple link-layers and tunnels in various combinations, some of which require decryption, such as:

> PPP over ATM
> Ethernet over ATM
> PPP over Ethernet over ATM
> IPsec over IP over ATM (or MPLS)
> PPP over L2TP over UDP over IPsec over IP over ATM (or MPLS)

In some cases multiple MPLS labels are stacked on a packet, all of which need to be removed as their tunnels are terminated.

Stage 2 - Identify Ingress Subscriber. This stage needs to be performed in connection with Stage 1, in cases where the link-layer protocols or tunnel headers provide information as to which subscriber this packet actually belongs, or where the subscriber identification provides the right keys for traffic decryption. Examples of such cases include PPP username authentication and IPsec, respectively. Other information in the packet, or about it (such as where in the system it originated), can also be used for ingress subscriber identification.

This multi-tiered approach to subscriber identification is necessary to support varying service provider business models, including wholesale and retail relationships. It is also the key to enabling end-user self-provisioning for advanced services on-demand, and supporting advanced services for mobile users.

Stage 3 - Filtering. Once the ingress subscriber is identified, filters can be applied according to customized policies specified by (or for) that subscriber. Filters can be set to either *deny* or *permit* traffic flow, and are applied against each incoming packet in a specific order. Filters can be designed to match on various attributes of IP packet headers - as well as on upper layer protocol headers - and on business application types, to enable stateful, dynamic filtering as required for firewall behavior. Robust filters can also allow concurrent enforcement of IPsec policies.

Stage 4 - Traffic Classification. Traffic classifiers allow the subscriber to have different traffic management, QoS, security and routing policies applied to different types of flows. Often these are used to distinguish

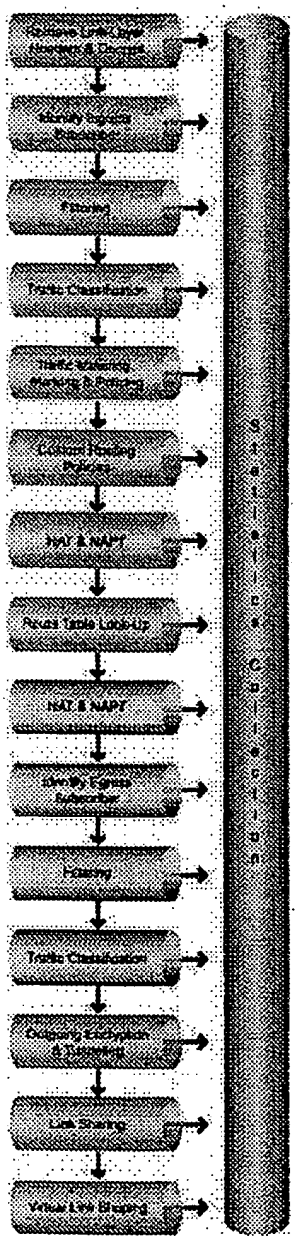

Figure 1: Packet Processing Stages for Advanced IP Services

different types of applications, such as latency-sensitive Voice over IP traffic vs. bulk data, but they can also be used to meet other business needs related to security or operational efficiency.

Stage 5 - Traffic Metering, Marking & Policing. Traffic meters are applied to each traffic class to control committed and peak information rates (CIR and PIR respectively), and their associated allowed burst sizes. Each traffic class of each subscriber can have its own meter with customized PIR and CIR settings. Traffic within the CIR limit is marked with the DiffServ Code Point (DSCP) for the intended Per Hop Behavior (PHB) for this traffic class. In the DiffServ protocol there are 64 possible PHBs or packet forwarding rules that define how to handle a given IP packet with respect to rules, such as drop precedence, weight and priority. Traffic in excess of the CIR and PIR limits is either dropped or assigned a lower grade PHB. This allows bandwidth on-demand, and prevents subscribers from using more than they are willing to buy at any given service level, while improving overall network efficiency.

Stage 6 - Custom Routing Policies. Some or all of the traffic from a subscriber can be directed to a) the Internet, b) a simple Virtual Private Network (VPN) tunnel to a pre-specified destination, or c) a Virtual Private Routed Network (VPRN). VPRNs allow subscribers belonging to the same customer organization to communicate in privacy (fully encrypted, if desired) over their own routed network, using their own private address space. In addition, traffic from suspicious locations or particular applications can be directed to an explicit next-hop address for special processing. This is useful for e-mail and virus screening, web caching and other advanced services.

"...marking traffic for a particular PHB is not the same as actually delivering that behavior."

Stage 7 - NAT & NAPT (Network Address [Port] Translation). If the subscriber is using a private address space, address translation services may be required to communicate with the Internet or another address space. Both static NAT with a one-to-one mapping of addresses, and dynamic, many-to-one NAPT mappings can be used. Address translation at this point is applied to the subscriber's source address.

Stage 8 - Route Table Look-up. The IP service switch now performs its most basic tasks - executing a best-matching-prefix look-up on the destination IP address and forwarding the packet to the correct next-hop host or router. This device must support the huge and growing size of the Internet routing table - nearly 100,000 entries - and run both interior and exterior routing protocols, such as OSPF and BGP-4. If subscriber policies dictate the use of a VPRN, the look-up must be done within the context of the correct private routing table.

HALFWAY THERE

At this point, it may seem that the processing is complete but it's not - it is only half done. More processing is required for the following reasons:

> The PHB determined by the ingress subscriber's traffic loads and policies must still be applied because marking traffic for a particular PHB is not the same as actually delivering that behavior.

> The packet may be destined for another customer, on the egress side of the same IP service switch, who has differing policies for QoS handling, traffic filtering and encryption.

> Further encryption or tunneling may be required to deliver the packet to its destination, and this may necessitate another route table look-up.

Continuing on, there are several additional processing stages that still must be performed to properly support egress subscribers and the final required elements of QoS enforcement.

Stage 9 – NAT & NAPT (Again). It is possible that the destination address belongs to a subscriber that requires address translation. By combining source and destination address translation, advanced services such as local- and wide-area load balancing and disaster recovery services can also be provided.

Stage 10 - Identify Egress Subscriber. If local to the same system as the ingress subscriber, the egress subscriber must now be identified and the appropriate policies must be applied as the packet is delivered.

Stage 11 - Filtering (Again). The local egress subscriber will likely have filtering policies on the packets being received, and these policies will also need to be applied by the IP service switch. As before, these can be based on dynamic, stateful filters specified at the application layer, or on other attributes of the IP and upper protocol headers.

Stage 12 - Traffic Classification (Again). Similarly, the egress subscriber may apply traffic classifiers, which can be different than those of the ingress subscriber, and override the QoS handling that was specified previously. One person's priority application may be another person's junk application.

Stage 13 - Outgoing Look-up, Encryption & Tunneling. Encryption may be required by either the ingress or the egress subscriber. If the packet is to be tunneled, a second route-table look-up may also be required. Appropriate link layer headers must also be applied.

Stage 14 - Link Sharing. Unavoidable congestion, such as that caused by port contention, should occur in only one place in a well-designed switch or router: at the output link. This allows proper QoS handling of each packet with the prescribed PHB. Combinations of Weighted Random Early Detection (WRED) with both Weighted Fair Queuing (WFQ) and priority-based queuing are powerful tools for creating a range of behaviors to handle all types of traffic appropriately in congested conditions. With up to 64 possible PHBs allowed by the DiffServ standard, this can require up to 64 queues per Virtual Circuit (VC), which can result in over 6.4 million queues for a system with 100,000 VCs - one for each subscriber.

Stage 15 - Virtual Link Shaping. Traffic sent on virtual backbone links, such as ATM VCs and MPLS tunnels or subscriber connections over any sort of technology, like Ethernet or POS, must often be shaped down to specific data rates. This egress shaping may define the boundaries within which all queues must share the link. Alternatively, different virtual links to the same next-hop destination can be used for different queues, each with a different capacity. This ensures that some PHBs are limited to a maximum capacity, thereby guaranteeing others a certain minimum capacity.

Stage 16 - Statistics Collection. Although mentioned last in this list, statistics collection must actually occur at all points during the flow of packets through the system. Detailed statistics must be efficiently gathered and made available for analysis at many levels, including per subscriber, per traffic class of each subscriber, per PHB, and per virtual link. Both real-time and historical data are necessary for management of customer Service Level Agreements (SLAs), as well as for accounting and billing purposes. Massive amounts of data must be collected and then rapidly reported out of the system.

Many existing architectures cannot support all of these processing stages and cannot meet stringent latency tolerances for real-time traffic. Common shortcuts eliminate or dramatically reduce support for IPsec, DiffServ and MPLS. Many routers oversimplify link sharing by only supporting a few queues and by shaping traffic only on ATM links. Most assume that traffic always flows from subscriber-to-backbone and back, and do not support subscriber-to-subscriber communication respecting both subscribers' policies. Likewise, the need for two IPsec operations, including a decryption and an encryption, on each packet at line speed, and with two distinct route table look-ups, is often overlooked. Finally, most routers cannot effectively collect the millions of statistics required for management and billing of the advanced IP services they are intended to offer.

IMPLEMENTATION ISSUES OF EARLY ARCHITECTURES

Performing the full range of complex processing tasks spelled-out above involves about a dozen table look-ups for each packet. Because payload handling is required for several of these stages, access to the full packet, not just the IP headers, is necessary. Additionally, many dozens of counters, flags and descriptors must be incremented, set and/or read as each packet flows through the system. Since there can be tens of thousands of concurrently active subscribers - each with different policies - comprising millions of different application flows, the size and numbers of the tables required can grow quite large. This poses significant challenges to the design of a high-performance solution.

"...most routers cannot effectively collect the millions of statistics required for management and billing..."

To support a subscriber with a fully loaded Gigabit Ethernet connection, for example, about 1.5 million packets per second must be processed. This means each packet must be processed through all the stages, in much less than a microsecond. Most modern general-purpose processors are optimized for high-speed number crunching rather than the processing of IP packets. They execute well from internal caches, but experience severe slowdowns when accessing external memory, regardless of internal clock speed. As such, their read and write speeds become a bottleneck for this sort of high-bandwidth operation. Since advanced packet processing requires many large separate tables, often with short entries, that are searched with short random-access queries, neither caching nor high-speed memory burst capability can help to increase processing throughput.

Combining several standard processors together in a cluster, and adding encryption or other specialized co-processors, can improve performance somewhat (up to about 30 Mbps with one encryption or decryption step and all other services deactivated, or to about 100 Mbps without encryption). Nevertheless, this rate remains far below that required for fast growing Internet PoPs that currently support thousands of subscribers. In an attempt to overcome this limitation, the first generation IP service routers used

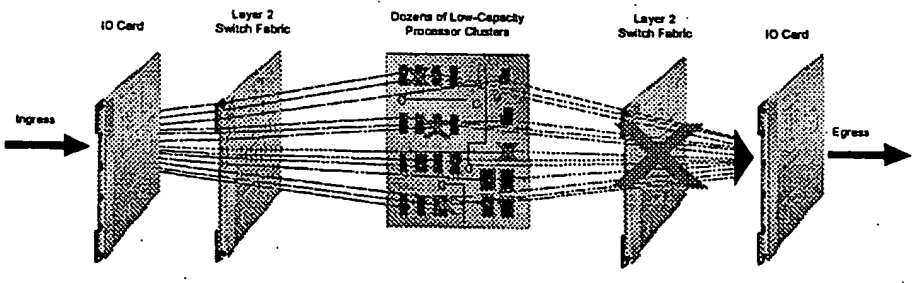

Figure 2: Congested Parallel Processing Architecture

parallel processing designs that employed many dozens of such processing clusters. (*See Figure 2.*) With this early design, incoming traffic is spread across the clusters using an intervening switch fabric, which directs the traffic from a single subscriber to a single cluster. Unfortunately, high-speed connections from individual subscribers still cannot be properly handled. If they exceed the throughput limitation of their assigned cluster, either their packets get dropped indiscriminately during the input stages, or certain advanced services are rendered unavailable to them. Furthermore, the system's processor resources must be managed carefully to protect against having too many subscribers assigned to the same processing cluster.

This early architecture requires each processing cluster to have high-speed access to huge routing tables, which also must contain subscriber-specific policies. In turn, this causes expensive duplication of table content and complex updating and synchronization procedures. If separate clusters are used for ingress vs. egress processing (to boost performance), then manual configuration of connections between these clusters is usually required, further complicating system management. Furthermore, this approach has an unpleasant side effect. It introduces unmanaged congestion, as traffic flows are re-aggregated on egress through the switch fabric, prior to reaching the output queue. Accordingly, the internal switch fabric of the router must decide which packets to drop, yet it has no concept of which packet belongs to which IP flow, or perhaps even which subscriber. Additionally, many switch fabrics have very shallow buffers, which eliminates the possibility to provide IP-layer QoS for latency or loss sensitive applications, and therefore can prevent the upholding of SLA commitments. To preserve QoS, systems based on this first generation architecture must be operated under light load and with little or no output link congestion. Scaling up the number of processing clusters only complicates operation by increasing the chance of unmanaged congestion on output. As a result, system costs grow faster than system performance.

"...systems based on this first generation architecture must be operated under light load and with little or no output link congestion."

THE NEXT GENERATION

A new architecture for delivering advanced IP services is now emerging in the form of a new class of product - the IP service edge switch. This next-generation platform solves the issues of performance and scalability using a hardware-based packet-processing pipeline.

This architecture distributes the load of table look-ups, packet manipulations and other functions along the pipeline, enabling large volumes of data to be processed at very high rates. The pipeline is constructed of specialized programmable network processors that are capable of performing various tasks on each packet at very high-speeds. The processors' packet-oriented flow-through architecture allows them to be cascaded together, easily extending the pipeline and therefore the system's overall processing capabilities.

Network processors alone, however, are not enough to handle the advanced packet processing required in an IP service edge switch. To deliver rich IP services at gigabit speeds and avoid bottlenecks, the most intensive processing stages in the pipeline must operate at least as fast as the rate of which packets arrive. To that end, the switch's pipelined processors must be complemented with high-speed custom ASICs and state-of-the-art encryption co-processors. ASICs, which still reign as the fastest way to perform most demanding packet processing tasks, handle the "heavy lifting" for specific tasks. This combination of ASICs and network processors provides an architecture that is flexible enough to address the evolving requirements of providing advanced IP services, while also delivering maximum performance. (*See Figure 3.*)

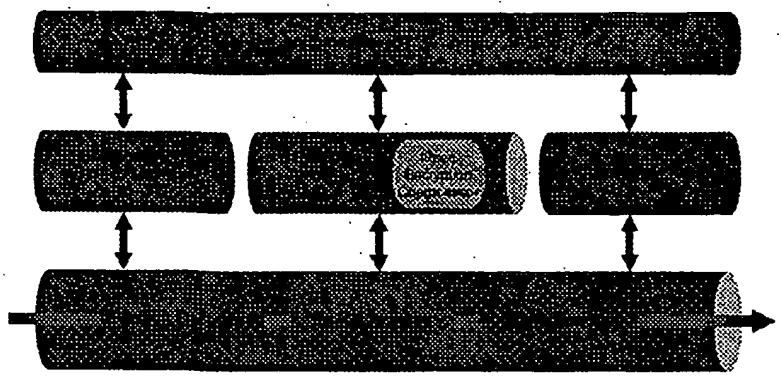

Figure 3: Fast Packet Processing Pipeline

An IP service edge switch with multi-gigabit capacity and redundancy can be built using several pipelines in parallel. Much simpler than the highly parallel architectures mentioned earlier, this approach scales far more easily and handles substantially larger individual data flows, since entire I/O ports up to Gigabit speeds are assigned to a dedicated pipeline. Unmanaged congestion losses are avoided by using an oversized back-end switching fabric among the pipelines themselves, resulting in a single pipeline feeding any given output link. With deep, fine-grained queues built into each pipeline, all subscriber QoS policies remain intact, even under extreme levels of output link congestion.

"With deep, fine-grained queues built into each pipeline, all subscriber QoS policies remain intact, even under extreme levels of output link congestion."

CHOOSING A SOLUTION

In evaluating IP service edge switches, NSPs should look for an architecture that allows a robust set of services to be rolled out over time. To take full advantage of the technology, the switch should provide the complete range of advanced IP services to every subscriber, including network-based VPNs, virtual routing, stateful firewall filtering, Optical-speed Encrypted Traffic Engineering, fine-grained QoS, NAT and NAPT, and more. Furthermore, it should provide advanced IP services concurrently with no performance degradation relative to basic IP routing, regardless of the number of subscribers and type of services enabled for each. And, the switch should easily support the development of additional revenues as new services are enabled, by providing detailed statistics for billing and SLA enforcement.

To achieve the high-touch optical-speed processing needed for delivering advanced IP services, a switching architecture like that described in this paper is required. Using a hardware-based pipeline built on specialized programmable network processors and custom ASICs will provide leading edge QoS, security, scalability and performance all at a relatively low price, while also maintaining the flexibility to adapt to new protocols and services.

Building a next-generation IP service edge switch that uses a hardware-based pipeline is a complex endeavor. Quarry Technologies, however, is uniquely qualified to meet this challenge. Designed by the same team that pioneered the separation of route calculation from forwarding and built the industry's first ASIC-based gigabit router, the Quarry Technologies' iQ4000™ and iQ8000™ IP Service Edge Switches implement all of the

advanced capabilities described above, via their pipelined architecture. As such, these products will accelerate the delivery and acceptance of advanced IP services. (*See Figure 4.*)

The Quarry Technologies iQ-series switches provide the foundation for customer-specific SLAs featuring application-level QoS concurrently with robust security. The switches support the industry's richest IP QoS implementation, together with gigabit-speed 3DES IPsec encryption. Remarkably, an encrypted packet flowing through the system can be decrypted, examined, classified, and then re-encrypted (under a different security association), as needed for routing traffic from one IPsec tunnel into another, all without any performance degradation. Up to 20 management and accounting statistics are collected per packet, and standard interfaces are supported to transfer this data to billing systems, so that all the value delivered by the switches may be billed accordingly.

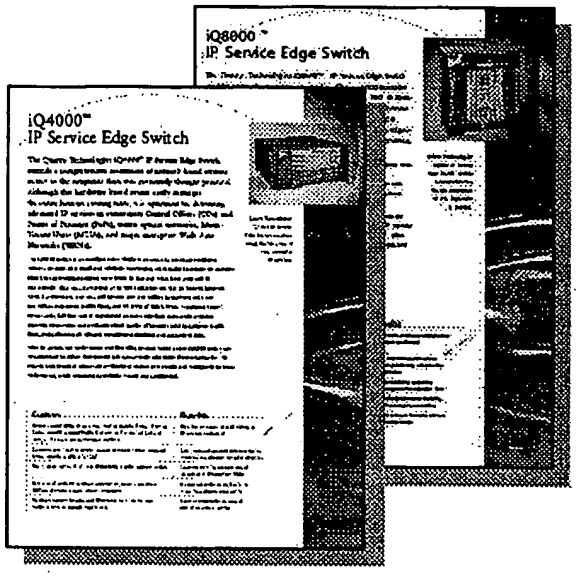

Figure 4: Quarry Technologies iQ-series IP Service Edge Swithces

In order for NSPs to rapidly and cost-effectively deploy high-margin, advanced IP services, the right hardware architecture is required. The new iQ-series IP Service Edge Switches from Quarry Technologies are built to provide intelligence for the service edge, making it easy for providers to deliver advanced IP services customized on-demand to the individual requirements of numerous diverse subscribers.

The introduction of the hardware-based iQ-series switches from Quarry Technologies will change the edge of the Internet. With this innovative, pipelined technology now available, it will no longer make sense to deploy equipment incapable of providing multiple, concurrent advanced services at high speeds. Instead, a new integrated and faster service edge will emerge. This will not only reduce costs and simplify operations and support, it will also speed the provisioning of new high-value services to all subscribers. In this way, NSPs will be able to capitalize on the investments they've made in increasing bandwidth, dramatically improving IP service delivery and thereby satisfying customers and increasing revenues.

Intelligence for the Service Edge

8 New England Executive Park, Burlington, MA 01803
Phone: 781.505.8300 Fax: 781.505.8316
Email: sales@quarrytech.com www.quarrytech.com

CORP-WP-700 070101